A.A. Ptukhin, A.E. Khrushkov, E.M. Bozhko
Ural Federal University named after the first President of Russia B.N. Yeltsin
Yekaterinburg, Russia

MACHINE LEARNING IN THE PROCESSING AND ANALYSIS OF TEXTS

Abstract: Natural language processing technologies have made great progress today, and considerable merit in this belongs to machine learning, which is used, particularly, for understanding texts. Neural network technologies can be used in any task where text classification is necessary, whether it is spam filtering, fraud defining or credit scoring, determining the mood of a text, or even the author’s tendency to be depressed, etc. In almost every paper in the collections of leading linguistic conferences, neural network methods are mentioned. Their popularity is largely due to their ability to find complex, sometimes hidden relationships in the data. However, in order for neural networks to fully demonstrate their practical effectiveness, large amounts of textual data are needed for training. This article tells about the language models used before the neural network revolution, whether it is possible to transfer the text to the computer's memory without losing its structure and semantics, and how a smartphone tells us words in messages, as well as about the use of neural network technologies in linguistics.

Keywords: machine learning, automatic text processing, text markup, neural networks.

А.А. Птухин, А.Е. Хрушков, Е.М. Божко
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина
Екатеринбург, Россия

МАШИННОЕ ОБУЧЕНИЕ В ОБРАБОТКЕ И АНАЛИЗЕ ТЕКСТОВ

Аннотация: Технологии обработки естественного языка сегодня шагнули далеко вперед, и немалая заслуга в этом принадлежит машинному обучению, применяемому, в частности, для понимания текстов. Нейросетевые технологии могут быть использованы в любой задаче, где необходимо классифицировать тексты, будь то фильтрация спама, определение мошенничества или кредитный скоринг, определение настроения текста или даже склонности автора текста к депрессии. Почти в каждой статье в сборниках ведущих лингвистических конференций используются нейросетевые методы. Популярность нейронных сетей во многом вызвана их способностью находить сложные, порой скрытые зависимости в данных. Однако для того, чтобы они могли в полной мере продемонстрировать свою практическую эффективность, необходимы большие объемы текстовых данных для эффективного обучения сетей. Данная статья рассказывает, какие языковые модели использовали до нейросетевой революции, возможно ли передать текст без потери структуры и семантики в память компьютера и каким образом смартфон подсказывает нам слова в сообщениях, а также о применении нейросетевых технологий в лингвистике.

Ключевые слова: машинное обучение, автоматическая обработка текста, разметка текста, нейронные сети.

The idea of artificial neural networks belongs to W. McCulloch and W. Pitts. In a joint study of the early 1940-ies, they proposed a formal model of the human brain – artificial neural network, and F. Rosenblatt later generalized their work and created a neural network model on the

computer. The use of textual data allows for effective training of networks and further use of them for automatic text processing.

Automatic text processing. The area of automatic text processing (ATP) emerged from the mix of computer (or mathematical) linguistics and machine learning. One of the main problems of the ATP is text presentation models: how to transfer the text to the computer memory so that its structure and semantics are preserved?

There are two fundamentally different approaches to this problem: a linear-algebraic vector model and a probabilistic language model. The first one is to a certain extent simpler: it implies that any document is represented by the frequency vector of the words contained therein. It is thanks to this model, which appeared in the mid-1970s and has not yet lost its relevance, that it became possible to use methods of traditional machine learning, such as a random forest method or a support vector method, and clustering algorithms – the k-means or hierarchical clustering methods [5].

A significant disadvantage of the vector model is that it does not allow to preserve the same word order. The language model was created precisely to solve it. It helps to answer two questions: what is the probability of a given sequence of words and which word is most likely to be seen after a given sequence of words? Before the neural network revolution, the parameters of the language model were, as a rule, estimated using the Markov chain apparatus. Markov language models are also imperfect: they can remember a small fixed number of previous words (for example, one or two), which, however, does not limit their applicability in such practical tasks as machine translation or speech synthesis.

Another problem that exists in the ATP is the representation of the meaning of the word. In traditional (cognitive) linguistics, it is believed that a person understands the meaning of words by context, that is, by the surrounding words. So, linguists came to the conclusion that the meaning of a word is represented by a vector of contexts showing which words could be encountered next to the given one.

Thus, in modern ATP, there are three structures: vector models of the document and words and language models – methods of both traditional and deep machine learning. The fundamental difference between traditional and in-depth machine learning is how to define the attribute space for to describe documents or words. In case of using traditional machine learning, it becomes necessary to use complex linguistic features that are often extracted from specialized knowledge bases, such as WordNet or FrameNet, or specialized (e.g. affective) dictionaries. It took

dozens of years to create such knowledge bases and dictionaries only for English, and in Russia these developments today are actively under way.

In case of in-depth training, specialized external resources are rarely attracted, and this significantly simplifies the work. However, since very large amounts of data are still needed to train the neural network, another problem arises, namely the markup of this data. Now virtually every particular task implies marking up new collections of documents. This process involves manual text processing: in each sentence the annotators need to highlight the words for the neural network to learn to extract. For a person, this task is simple and meaningful but time-consuming, so the markup is often conducted on crowdsourcing platforms, where you can attract for relatively small financial reward a lot of anonymous annotators that do not have a professional linguistic education.

Speech recognition technology. Today, most research or practical projects in the ATP uses the neural network technology (not necessarily a deep one) in one way or another. Particularly widespread are models of distributed word vectors allowing representing a word with such vectors, on which arithmetic operations, such as adding or subtracting, can be performed. A classic example is shown in the works of Th. Mikolov, where he demonstrated how the solution of the equation «king - man + woman = queen» using such vectors. These equations are also valid for morphological analogies, e.g.: «good - good + bad = bad». Since then, not only the complexity of the architectures used in the ATP tasks has increased many times, but alternative training paradigms have also been used, e.g., reinforcement learning, which allows to teach a neural network not to predict the correct values but to behave in a certain way (say, to generate texts of a particular genre or maintain long dialogues). Apparently, it is because of the lack of large markings that this interest in depth learning came to the ATP rather late as compared to computer vision: in 2003, a paper on neural language models by J. Benjo was published, and in 2011, the work by R. Collobert and J. Weston was issued. In both papers, relatively simple neural network architectures were used to solve standard tasks – predicting the next word in Benjo's work and determining the part of the word's speech and extracting named entities (names of country, company, and people) in Collobert and Weston's work. However, a significant leap was shown in the quality of their solution [2].

Machine learning in linguistics. Another architecture that shows standard results on the order of 90% correct answers in a variety of tasks,

from allocating named entities to machine translation, is a two-layer recurrent network with long short-term memory cells (biLSTM). It can read input sentences from right to left and from left to right and transfers the meaning of each individual word and the entire sentence into separate vectors. It is for these vectors that the neural network makes a decision whether, for example, the word is a city name or a surname, or learns to generate a translation into another language This architecture often uses the so-called attention mechanism: it turns out that when adding a very simple add-in to the recurrent layers, the neural network can be taught to focus on the important elements in the sentence. This mechanism is interesting for two reasons: firstly, despite its simplicity, it gives a tangible increase in quality indicators, and secondly, it helps to exactly understand how the neural network is learning and makes it more transparent.

Analysis features. There are different types of neural networks, but when working with textual data, two of them are used most often: convolutional and recurrent neural networks (recursive neural networks are used less often).

Convolution is a special kind of operator that assigns a scalar to several vectors. As a rule, several convolutions are used, and several scalars are obtained on the output.

Convolutional neural networks came to ATP from computer vision. The motivation for using the convolution layer there is quite natural: the image usually contains pixels similar in color, so there are areas of a certain color, the color of each particular pixel being irrelevant. When working with texts, the convolution layer is interpreted differently: it helps to find stable n-grams, i.e., sequences of *n* words that are significant to the problem being solved; e.g., for the problem of thematic classification, such n-grams may be terms, names, and stable collocations [1].

The idea of using n-grams as attributes has been widely used before, but, as with language models, neural network methods require explicit indication of the value *n*, and using convolutions allows to overcome this constraint by means of specifying *n* implicitly. Convolutional neural networks are mainly used in various classification tasks: thematic classification, definition of a paraphrase and extraction of relationships between words. In the first case, for a pair of sentences, the neural network must decide whether their meaning is the same (e.g., if the sentences *«The son of Mary is 18 years old»* and «*He was born in 2000, and his mother is Maria»* are the same in meaning), and in the second case the neural network has to determine whether a pair of objects is in a particular

relationship (e.g., whether the words *«engine»* and «*machine»* in the sentence: *«After washing the car, the engine stopped starting»* are in the «part-whole» relation).

Convolutional neural networks. If convolutional neural networks allow to find short dependencies between words, then recurrent neural networks are created in order to take long dependencies (e.g., within the same sentence) into account. The recurrent network reads the sentence and calculates vectors representing each word separately and the sentence as a whole [3].

Until recently, these architectures had rarely been used for texts. The reason is quite simple: the text is a collection of discrete features, that is, words or even letters. Despite the fact that each word or letter can be represented by a vector consisting of numbers, these numbers cannot be slightly changed to obtain another word or another letter. Over the past couple of years, many research projects dedicated to the adaptation of these architectures for working with textual data have emerged.

Computational linguistics. When using in-depth training in ATP, the problem of transfer of training is particularly acute: a neural network trained to solve a specific task on one body of texts may not cope with the same task on another body. Unfortunately, the neural networks trained in good academic buildings can show very poor results in practice, and it is not due to their bad architecture or ideology itself, but because the real user texts are far from those used in network training [5].

Nevertheless, we can certainly say that the use of neural network technologies has greatly improved the quality of solving many problems, in particular, the problems of determining the part of speech and parsing the sentence, and led to the creation of new research directions. So, now different schools of machine translation are distinguished: statistical machine translation and neural network translation.

REFERENCES

1. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М.: МИЭМ, 2011. 272 с.

2. Kyunghyun Cho. Introduction to Neural Machine Translation with GPUs (part 3). // NVIDIA Developer Blog [Electronic recource]. URL: https://devblogs.nvidia.com/introduction-neural-machine-translation-gpus-part-3/ (22.12.2018).

3. Старостин А. ScienceHub #06: Компьютерная лингвистика // Хабр [Электронный ресурс]. URL: https://habr.com/ru/company/abbyy/blog/202306/ (дата обращения: 15.12.2018).

4. Тарасов Д. Классификация предложений с помощью нейронных сетей без предварительной обработки. // Хабр [Электронный ресурс]. URL: https://habr.com/ru/company/meanotek/blog/256593/ (дата обращения: 23.12.2018).

5. Черняк Е. Глубинное обучение в обработке и анализе текстов. // Postnauka.ru [Электронный ресурс]. URL: https://postnauka.ru/longreads/85951 (дата обращения: 30.11.2018).